图书情报工作
第 64 卷 第 12 期　2020 年 6 月

多维度疾病语义相似度研究*

■ 张军亮[1,2,3]

[1] 新乡医学院管理学院　新乡 453003　[2] 新乡医学院卫生信息资源研究中心　新乡 453003

[3] 河南省健康中原研究院　新乡 453003

摘　要：[目的/意义] 针对疾病知识的不同表达方式，提出一种融合疾病多维度的综合语义相似度计算方案。[方法/过程] 在整合疾病本体和医学百科各自特征的基础上，设计由基于疾病本体的语义相似度和基于医学百科的疾病语义相似度构成的综合语义相似度模型。其中，运用图论计算基于疾病本体的语义相似度，运用 LDA、集合和向量空间模型计算基于医学百科的疾病语义相似度。[结果/结论] 将本文的方法同临床医生的人工判别进行比较，结果表明本文的方法能够有效地反映疾病的语义相似度。本文的方法可为疾病相似性进一步研究提供参考。

关键词：语义相似度　疾病本体　疾病百科　相似度计算
分类号：G254
DOI：10.13266/j.issn.0252-3116.2020.12.14

1　引言

语义相似度被用于反映概念术语或者文档间的相似程度，一直是人工智能、认知科学、自然语言处理的热点和难点[1]。语义相似度在信息检索、服务推荐、文本聚类分析等方面有广泛的应用[2-4]。疾病语义相似度对研究疾病的发病机制、诊断和药物研制具有积极的作用，被广泛地应用于生物医学概念术语关系的研究中[5-6]。

在对医学信息进行知识管理的过程中，研究者从不同方面建立了多种包含医学概念关系的知识库，如：由世界卫生组织依据疾病的病因、病理、临床表现和解剖位置等建立的国际疾病分类表（International Classification of Diseases，ICD）和由美国国立医学图书馆建立的医学主题词表（Medical Subject Headings，MeSH）等。近年来，研究者运用本体来描述医学知识间的关系，建立了一些医学本体知识库，来实现生物医学术语的语义可计算性。生物医学计算中心的 BioPortal[7] 整理统计有近一千个医学本体。L. Schriml[8] 为了实现人类疾病的形式化表示，利用本体技术构建了疾病本体。国内也积极展开医学知识组织相关研究，朱玲等[9] 以我国中医学文献为基础，进行中医本体构建研究；李兰娟院士团队构建了肝炎本体[10]。

随着医学科学的发展，人们积累了大量的医学知识。医学百科全书是一种重要的知识表示形式，国内外专家出版了大量的医学百科全书，如：由全球数百位医学专家、一个独立的同行评审编辑委员会和专业医学作者协作撰写而成《默沙东诊疗手册》和由我国政府主导完成的《中华医学百科全书》。互联网时代，促进了百科全书的发展，形成了 Wiki 百科和百度百科等互联网百科，在此基础上也形成了一批医学互联网百科，如：美国国立医学图书馆建立 Medlineplus 和我国卫健委牵头建立百科名医中的医学百科。

如何运用不同类型的医学信息资源来计算疾病的语义相似度，提高疾病语义相似度计算的全面性和准确性，将有利于医学信息资源的发现服务，为更深层次的智慧医学发展提供支撑。基于此，本文研究利用疾病本体与疾病百科全书中关于疾病描述的信息，设计了基于多维度的疾病语义相似度的计算方法，首先分析国内外语义相似度研究中常用的计算方法；然后，研究融合疾病本体和百科全书的多维度的疾病语义相似度的计算方法；最后，利用具体实例对本文提出的方法进行分析。

2　相关研究

语义相似度受到研究者的广泛关注，根据研究对

* 本文系国家社会科学基金青年项目“基于语义关联的多源医学信息资源发现服务体系研究”（项目编号：17CTQ026）研究成果之一。
作者简介：张军亮（ORCID：0000-0002-3678-8691），副教授，博士，新乡医学院太行青年学者，E-mail：junliangzhang2000@163.com。
收稿日期：2019-05-13　修回日期：2019-08-22　本文起止页码：127-135　本文责任编辑：杜杏叶

象和任务的差异,语义相似度可以分成概念(词语)层次和文本(句子、段落)层次[11]。概念语义相似度是对词语间的关系进行语义度量[12]。S. Spagnola 等[13]运用概念在语义网中的最短路径,融合用户评分等特征来表示语义相似性。R. Cilibrasi 等[14]以万维网作为数据库,以 Google 搜索引擎为基础,构建了谷歌语义相似度的计算方法。也有学者利用现有的语义知识库中的语义关联计算概念语义相似度,李峰[15]和刘杰[16]分别以 HowNet-2000 和 HowNet-2008 为基础研究中文概念的语义相似度;T. Nguyen[17]和 X. Liu[18]利用 WordNet 计算词语间的语义相似度。张军亮等[19]运用农业百科中词条注释来计算语义相似度。文本语义相似度是计算句子或段落间语义相关程度的[20]。I. Aainul 等[21]运用语料库和最长公共子序列来研究句子或段落间的语义相似度。Q. Chen 等[22]运用 LDA 来计算短文本的语义相似度。M. Farouk[23]利用词嵌入向量和 WordNet 来计算两个句子间的语义相似度。李琳等[24]利用依存句法分析和词嵌入向量相结合的方法计算句子间的语义相似度。詹志建等[25]在利用复杂网络表征短文本的基础上计算短文本的语义相似度。

依据语义相似度的实现算法,语义相似度可以分成基于统计方法、基于图论方法和基于混合技术方法[26]。基于统计方法主要是在语料库的基础上,运用词语共现、上下文信息等对概念或文本进行表示,再结合数学运算来计算语义相似度。D. Bollegala 等[27]利用 Web 搜索引擎返回的页面计数和文本片段来计算语义相似度;基于图论方法是在现有知识库的基础上,运用图论的相关理论来解释语义相似度[28]。R. Rada 等[29]利用两个概念间的最短路径来度量概念语义相似度。A. Banu 等[30]将概念所包含的子概念也作为语义概念的影响要素。X. Zhu 等[31]将图局部区域密度引入到语义相似度计算中,来改进相似度的效果。李文清等[32]将信息论理论引入到概念语义相似性计算中,提出了一种加权本体概念语义相似度计算方法;基于混合技术方法是针对多源信息综合运用多种方法进行语义相似度计算的方法。L. Sahni 等[33]整合 Web 搜索引擎的相似性度量和词语的分类结构相似性度量来实现语义相似度的计算。Y. Yang 等[34]综合概念间的语义距离、概念层次以及上下义词集合之间的重叠程度来量化概念间的语义相似性。

语义相似性也是生物医学研究过程中的重要内容,如基因聚类、基因表达数据分析、分子相互作用的预测等,生物医学领域的语义相似度主要基于现有的本体和医学知识库展开[35-36]。J. Jeong 等[37]和 P. Dutta 等[38]利用基因本体研究基因和基因产物的语义相似度。H. Al-Mubaid 等[39]在 UMLS 框架下,使用 Medline 作为标准语料库和网格本体来测量生物医学领域概念之间语义相似性的可行性。李文庆[40]运用比较概念的所有分类知识方法,提出了一种医学语义相似度算法。

综上所述,现有语义相似度概念计算方法主要是基于本体知识的概念层级关系和运用文本相似度来实现,但也存在一些问题:①较少针对同一语义概念将两者整合起来实现综合语义相似度计算。将不同的知识资源采用不同语义计算方法,并有效地整合起来计算语义相似度,可以全面反映概念间的语义相似度;②文本相似度计算中,一般将概念的描述文本作为一个整体文本进行分析,较少关注针对概念的不同描述表达。采用不同的计算方法,将描述概念的文本依照内容类别进行分解,并针对文本描述的不同进行分别处理,可以更科学合理地反映概念的语义相似。

由于疾病语义相似度研究具有重要意义,并且疾病有本体和医学百科等多种不同的知识表达形式,同时,医学百科中疾病的概念由概述、症状、病因、诊断和治疗等多部分组成,因此,本文整合疾病本体和医学百科信息资源,将医学百科中疾病概念的不同描述分别处理,设计了基于疾病的知识表达和内容描述的多维度疾病语义相似度计算方法,以提高疾病语义计算的全面性和合理性。

3 疾病综合语义相似度计算

疾病概念的表示形式有多种多样,结合不同表示形式的特点,选择相似度计算方法,并将不同语义相似度有效整合起来,可以得到更全面、更准确的语义相似度。本文将来自疾病本体和医学百科中疾病词条的内容整合起来,对疾病的语义相似度进行分析,具体方案见图1。疾病综合语义相似度的计算过程如下:

首先,把两个疾病词语在疾病本体中找到对应的疾病概念,基于疾病本体(Disease Ontology, DO)计算疾病间的语义相似度 $So(w_1, w_2)$。

然后,在医学百科中查找到两个疾病对应的词条,利用医学百科内容的相似度来计算疾病间的语义相似度 $Sd(w_1, w_2)$,需要通过计算定义语义相似度、症状语义相似度、病因语义相似度、诊断语义相似度和治疗语义相似度得到相关数据。

最后,将基于DO的疾病语义相似度和基于医学百科的疾病语义相似度通过公式(1)综合起来。

$$S(w_1,w_2)=\alpha*So(w_1,w_2)+\beta*Sd(w_1,w_2)$$

公式(1)

其中 $\alpha+\beta=1$,依据需求调节两个相似度之间的比例,设置的基本原则是使综合疾病的综合语义相似度与工判读的疾病语义相似度尽可能一致。

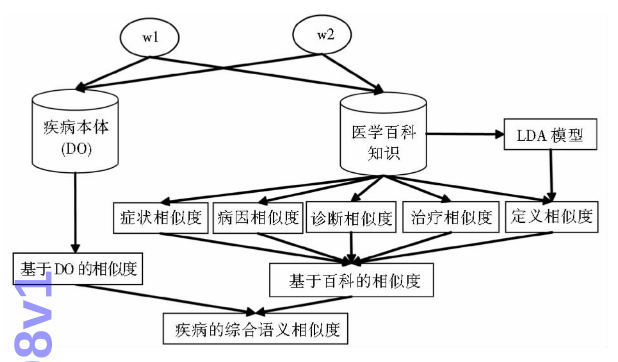


图1　疾病综合语义相似度计算方案

3.1　基于DO的疾病语义相似度计算

DO将每个疾病概念作为一个节点,通过概念语义关联,建立的本体知识库,并且同MeSH、ICD、SNOMED和OMIM知识库中的疾病概念术语联系起来。图2是DO中代谢疾病类的部分结构:

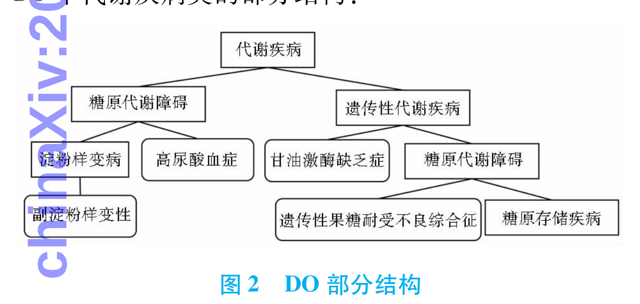


图2　DO部分结构

参考J. Zhang[41]文献中关于概念树形结构相似度的计算方法,疾病本体的语义相似度通过公式(2)来计算:

$$So(w1,w2)=\frac{depth(NCW(w_1,w_2))}{depth(w_1)+depth(w_2)-depth(NCW(w_1,w_2))}$$

公式(2)

$So(w_1,w_2)$的值为[0,1],值越大表明在疾病本体中,两个疾病概念越相似。

定义1:$depth(w)$表示概念w到根节点的深度,即节点到根节点的距离。如在图2中,“副淀粉样变性”距离根节点的“代谢疾病”的距离为3,$depth$(副淀粉样变性)=3。

定义2:$NCW(w_1,w_2)$表示距离概念 w_1 和概念 w_2 最近的共同祖先概念。如在图2中,“遗传性果糖耐受不良综合征”(w_1)和“甘油激酶缺乏症”(w_2)最近的共同祖先节点为“遗传性代谢疾病”,$NCW(w_1,w_2)$ = 遗传性代谢疾病。

如在图2中,“遗传性果糖耐受不良综合征”(w_1)和“甘油激酶缺乏症”(w_2)的语义相似度计算为 $\frac{1}{2+3-1}=0.25$。

3.2　基于医学百科的疾病语义相似度计算

在医学百科中,疾病条目对每个疾病都从概述、症状(包含临床表现等)、病因、诊断(包含检验等)和治疗(包含预防等)等对疾病知识进行了比较完备的解说。通过对条目分析,疾病概述部分浓缩了疾病的本质基本知识;症状部分描述了疾病的表现症状;病因部分详细说明了疾病的发病原因;诊断部分详细描述了疾病诊断的过程;治疗部分详细阐述了疾病治疗方案等。

由于疾病条目的概述、症状、病因、诊断和治疗部分对于疾病的描述表示在文本的长度方面存在差别,且各部分的描述语言和词语的语义密度也存在差异,如:概述部分内容相对较短,语义密度相对较集中;症状部分医学术语描述相对较多;病因、诊断和治疗部分内容相对较长。因此,本文针对各部分的描述特征的差异,设计了不同的语义相似度计算方法。疾病概述部分的词语相对较少,且词语之间存在较高的关联性,LDA可以识别文本中潜藏的主题信息,因此疾病概述部分设计了基于LDA的相似度计算方法;症状部分的词语多是疾病的临床表现等病人的异常感觉或某些客观病态改变,可以理解为词语的集合,因此,症状部分设计了基于集合的相似度计算方法;病因、诊断和治疗部分与其他一般文本内容相似,但是各部分的词语出现的频次等相关特征存在差异,因此病因、诊断和治疗部分分别设计了基于向量空间的相似度计算方法。

3.2.1　基于LDA的相似度计算

2003年D. Blei等依据词的共现,结合“word-document-topic”提出隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)模型[42]。LDA作为一种非监督的机器学习方法,被广泛地应用到文本信息分析中[43]。本文将医学百科中的疾病词条的概述作为疾病定义,运用LDA模型得到每个疾病的主题分布,由于主题模型是以概率的形式存在,因此,相似度计算采用相对熵[44]来计算。疾病定义相似度计算的步骤:

(1)对百科中的疾病词条的概述部分进行分词,

并抽取其中的医学术语和名词,形成数据集;

(2)利用 LDA 模型算法对训练数据集进行分析处理,得到"topic-word"的 LDA 模型;

(3)对利用 LDA 模型计算疾病 w 的主题分布 Tw,利用定义 3 计算疾病的相似度。

定义 3:Tw 为疾病定义的主题分布,$(t_1, t_2, \cdots, t_n)$,疾病 w1 和疾病 w2 的定义相似度为:

$$Sde(w_1,w_2) = 1 + \sum_{i=1}^{n} \frac{t_{1i}+t_{2i}}{2} \ln \frac{t_{1i}+t_{2i}}{2} - \frac{1}{2}\left(\sum_{i=1}^{n} t_{1i} \ln t_{1i} + \sum_{i=1}^{n} t_{2i} \ln t_{2i} \right) \quad 公式(3)$$

其中,n 为 LDA 模型中主题数量,t_{1i} 和 t_{2i} 分别为 w_1 和 w_2 的第 i 主题的概率,$Sde(w_1,w_2)$ 的值为[0,1],值越大表明两个疾病概念越相似。

3.2.2 基于集合的相似度计算

疾病的症状相似度通过计算描述不同疾病的症状来进行表示,具体通过定义 4 来计算。

定义 4:疾病 w 症状由 $set(w)$ 表示,即描述疾病症状的术语集合,疾病 w_1 和疾病 w_2 的症状相似度为:

$$ssy(w_1,w_2) = \frac{set(w_1) \cap set(w_2)}{set(w_1) \cup set(w_2)} \quad 公式(4)$$

其中 ∩ 为两个集合交运算,∪ 为两个集合的并运算,$Ssy(w_1,w_2)$ 的值为[0,1],值越大表明两个疾病具有相同症状越多,两个疾病概念越相似。

3.2.3 基于向量空间的相似度计算

向量空间模型是对文本内容进行向量化处理,使其可以进行向量化处理,并且能够以空间上的相似度表示语义相似度,被广泛地应用于文本信息处理中。疾病百科中关于病因、诊断和治疗的内容相对较丰富,因此本文采用基于向量空间的相似度计算方法。具体的实现步骤为:首先对病因、诊断和治疗部分的文本向量化;然后分别利用定义 5、定义 6、定义 7 计算相似度;最后,通过定义 8 计算疾病基于医学百科的疾病相似度。

定义 5:疾病 w 病因的词向量定义为 ws,疾病 w_1 和疾病 w_2 的病因相似度为:

$$Set(w_1,w_2) = \frac{ws_1 \cdot ws_2}{\| ws_1 \| * \| ws_2 \|} \quad 公式(5)$$

其中 ws_1 和 ws_2 为疾病 w_1 和疾病 w_2 病因部分的文本向量,· 为向量内积,‖ 为向量模运算。

定义 6:疾病 w 诊断的词向量定义为 wd,疾病 w_1 和疾病 w_2 的诊断相似度为:

$$Sdi(w_1,w_2) = \frac{wd_1 \cdot wd_2}{\| wd_1 \| * \| wd_2 \|} \quad 公式(6)$$

其中 wd_1 和 wd_2 为疾病 w_1 和疾病 w_2 诊断部分的文本向量。

定义 7:疾病 w 治疗的词向量定义为 wt,疾病 w_1 和疾病 w_2 的治疗相似度为:

$$Str(w_1,w_2) = \frac{wt_1 \cdot wt_2}{\| wt_1 \| * \| wt_2 \|} \quad 公式(7)$$

其中 wt_1 和 wt_2 位疾病 w_1 和疾病 w_2 治疗部分的文本向量。

$Set(w_1,w_2)$、$Sdi(w_1,w_2)$、$Str(w_1,w_2)$ 的值为[0,1],值越大表明两个疾病在病因、诊断和治疗方面具有相似性越大。

定义 8:基于疾病描述的语义相似度为:

$$Sd(w_1,w_2) = \gamma_1 * Set(w_1,w_2) + \gamma_2 * Ssy(w_1,w_2) + \gamma_3 * Set(w_1,w_2) + \gamma_4 * Sdi(w_1,w_2) + \gamma_5 * Str(w_1,w_2) \quad 公式(8)$$

其中 γ_1、γ_2、γ_3、γ_4、γ_5 各个语义相似度的权重,且 $\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 + \gamma_5 = 1$,通过专家和实验对其调整设置,设置的基本原则是依据内容反映疾病语义的程度来设定,概述部分是对疾病相对全面的概括,设置较高的权重;病因、症状、诊断和治疗描述疾病语义的不同方面内容,对疾病语义相似度影响程度认为是均等的。

4 实验

为了验证本文提出的疾病语义相似度方法的效果,利用本文方法对 20 对疾病进行了相似度计算,同时组织临床医生对 20 对疾病进行了相似性判别,对本文方法和人工判别进行相关性比较。

4.1 实验环境

本文的数据集主要包括疾病本体和疾病百科。疾病本体数据来源于西北大学基因医学中心和马里兰大学医学院的基因组科学研究所的 Disease Ontology[45]。疾病概念描述利用国家卫健委建设的百科名医网中疾病百科[46],共收集7 808个疾病概念。本文利用百科名医网中疾病百科药品百科和检验百科中的医学词语,以及搜集了医学症状的词语组成医学词典,将医学词典作为疾病概念描述的分词词典。

临床医生是从事临床治疗和医学研究的一线,最能掌握疾病间的语义关联,因此本文还组织了新乡医学院第一附属医院(三级甲等)的临床医生对疾病相似度进行判断,从内分泌科和神经内科两个科室选择 5 位临床医生(其中 1 位主任医师、2 位副主任医师和 2 位主治医师)参与对 20 对疾病的相似度进行独立人工评判,相似度等级为 0 到 9,0 表示完全不相似,9 表

示完全相似,最后利用公式(9)计算人工相似度:

$$sp = \frac{\sum_{i=1}^{5} sp_i}{5 * 9} \qquad 公式(9)$$

其中 sp_i 为第 i 个专家对疾病对的相似度评判分数。sp 的值为[0,1],0 表示临床医生认为两个疾病概念完全不相似,1 表示临床医生认为两个疾病概念完全相似,sp 值越大表明临床医生认为两个疾病的相似度越高。为了验证五位临床医生在疾病语义判断的一致性,对其结果运用克朗巴赫系数(Cronbach's Alpha)[51]进行了一致性检验,结果为 0.977,表明医生之间的疾病语义相似度判断具有较高的一致性。

表 1　疾病语义相似度

疾病 1	疾病 2	人工	综合 1	综合 2	综合 3	本体	定义	病因	症状	诊断	治疗	综合 4
2 型糖尿病	1 型糖尿病	0.67	0.55	0.51	0.58	0.71	0.61	0.14	0.12	0.08	0.55	0.38
2 型糖尿病	百日咳	0.00	0.03	0.03	0.02	0.00	0.11	0.01	0.04	0.01	0.02	0.06
2 型糖尿病	糖尿病	0.78	0.71	0.69	0.73	0.80	0.82	0.65	0.17	0.58	0.57	0.62
2 型糖尿病	妊娠糖尿病	0.67	0.57	0.55	0.59	0.67	0.75	0.17	0.00	0.57	0.47	0.48
2 型糖尿病	糖尿病视网膜病变	0.49	0.17	0.20	0.13	0.00	0.80	0.02	0.00	0.05	0.04	0.34
2 型糖尿病	先天性心脏病	0.00	0.04	0.04	0.03	0.00	0.14	0.05	0.02	0.01	0.03	0.07
2 型糖尿病	高血糖	0.51	0.45	0.42	0.47	0.57	0.47	0.09	0.10	0.66	0.05	0.32
2 型糖尿病	低钾血症	0.18	0.18	0.17	0.20	0.25	0.27	0.02	0.02	0.01	0.02	0.12
病毒性肺炎	风湿性关节炎	0.22	0.13	0.14	0.12	0.08	0.39	0.02	0.03	0.06	0.01	0.17
病毒性脑膜炎	埃博拉出血热	0.20	0.17	0.21	0.14	0.00	0.51	0.69	0.05	0.07	0.10	0.34
病毒性脑膜炎	病毒性肺炎	0.29	0.22	0.24	0.20	0.10	0.47	0.54	0.01	0.20	0.25	0.34
病毒性脑膜炎	风湿性关节炎	0.13	0.08	0.08	0.08	0.09	0.09	0.01	0.07	0.10	0.03	0.07
病毒性脑膜炎	白癜风	0.00	0.08	0.08	0.09	0.11	0.10	0.00	0.07	0.02	0.01	0.05
病毒性脑膜炎	鼻炎	0.18	0.15	0.16	0.14	0.10	0.42	0.14	0.02	0.03	0.03	0.20
风湿性关节炎	白癜风	0.04	0.15	0.16	0.14	0.09	0.47	0.04	0.05	0.01	0.01	0.20
风湿性关节炎	鼻炎	0.18	0.11	0.12	0.11	0.08	0.29	0.03	0.07	0.04	0.03	0.14
埃博拉出血热	病毒性肺炎	0.20	0.18	0.22	0.14	0.00	0.63	0.42	0.06	0.14	0.10	0.36
埃博拉出血热	风湿性关节炎	0.00	0.03	0.04	0.02	0.00	0.13	0.01	0.02	0.01	0.01	0.06
埃博拉出血热	白癜风	0.00	0.03	0.04	0.03	0.00	0.15	0.00	0.01	0.02	0.01	0.07
病毒性肺炎	鼻炎	0.33	0.25	0.26	0.24	0.18	0.67	0.11	0.07	0.08	0.05	0.32

注:综合 1 相似度 $\alpha = 0.5, \beta = 0.5$;综合 2 相似度 $\alpha = 0.6, \beta = 0.4$;综合 3 相似度 $\alpha = 0.4, \beta = 0.6$;综合 4 为百科综合相似度 $\gamma_1 = 0.4, \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = 0.15$

本文在实验过程中使用的编程语言环境 Python3.6 64 位系统[47],自然语言处理工具为 HanLP[48],数学计算 Numpy[49],主题分析 gensim[50]的 LDA 和 TF-IDF。

4.2　评价方法

为评价本文提出算法的有效性,采用 Spearman 相关系数和 Pearson 相关系数来进行评价。两个具有相同数量元素的随机变量 X(疾病的综合语义相似度)、Y(人工判读的疾病语义相似度),X_i、Y_i 分别为 X、Y 中的第 i 个元素,对 X、Y 中的元素按照升序或降序的方式对其进行排序,x_i、y_i 分别为 X_i、Y_i 的排序位置,将集合 X、Y 中的对应元素的位置进行差运算得到 $d_i = x_i - y_i$,Spearman 相关系数[51]计算公式为:

$$p = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2 - 1)} \qquad 公式(10)$$

其值在[－1,1]之间,其值越大,表示其相关性越大。本文利用 Spearman 相关系数反映本文的疾病综合语义相似度计算方法与临床医生对疾病相似度之间的相关关系。如果 Spearman 相关系数越接近 1,表明本文设计的疾病语义相似度与医生的认知判断越相近。

Pearson 相关系数[51]计算公式为:

$$p = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N} X_i^2}\ \sqrt{\sum_{i=1}^{N} Y_i^2}} \qquad 公式(11)$$

其值也在[－1,1]之间,Pearson 相关系数的绝对值越大,相关性越强,相关系数越接近于 1 或－1,相关度越强,相关系数越接近于 0,相关度越弱。同 Spearman 相关系系数一样,如果 Pearson 相关系数越接近 1,表明本文设计的疾病语义相似度与医生的认知判断越相近。

4.3 实验过程及结果

4.3.1 实验过程

在实验中,包括计算基于疾病本体的语义相似度和基于疾病百科的语义相似度两部分,基于本体的语义相似度利用疾病到根节点的路径来实现计算;基于疾病百科的语义相似度需要先计算概述的 LDA 模型,病因、诊断和治疗的 TF-IDF 模型,模型实现过程如图 3 所示:

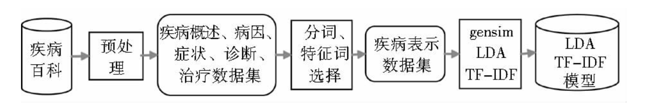


图 3 疾病百科的 LDA、TF-IDF 模型建立过程

实验中预处理的主要工作是从网络上将医学百科的疾病词条相关内容收集起来,综合利用内容的结构将疾病条依照疾病概述、病因、症状、诊断和治疗部分进行分割处理,实现对内容的清洗,为下一步各部分的数据集构建提供原始材料。

实验中分词、特征词选择过程,首先,将医药相关术语加入到医学,并将其词性标注为"nh",同时加入汉语的停用词词典;然后,利用 HanLP 的分词工具分别对疾病的不同部分进行分词处理;最后,依据分词的词性选择不同的特征词,其中概述、病因、诊断和治疗选择名词性词语,症状部分主要选择词性为"nh"的医药相关词。

实验中 TF-IDF 模型的实现过程,分别利用疾病的病因、诊断和治疗的特征词,构建各自的语料库,然后利用 gensim 中 TfidfModel 模块构建各自的 TF-IDF 模型。

实验中 LDA 模型的实现过程,利用疾病概述部分的特征词,构建语料库,利用 gensim 中的 LdaModel 构建 LDA 模型。

在疾病概述的 LDA 模型建立中,主题数的确定对于模型的应用至关重要,本文在实验中利用困惑度[52]来确定主题数,不同主题数的困惑度见图 4。

通过图 4 可以得到主题数设置为 60,迭代数设定为 1 000 次,LDA 模型的困惑度最小,因此,本文在 LDA 模型中主题数选择为 60,迭代次数设定为 1 000。

实验中基于本体的语义相似度过程,首先是获得聚类两个概念最近的共同的节点,然后,分别获取三个节点到根节点的距离,最后,利用公式(2)计算相似度;基于医学百科的疾病语义相似度计算,基于 LDA 的相似度过程,首先,导入 LdaModel 模型,然后利用疾病的概述分词和特征提取的特征词和 LdaModel 模型

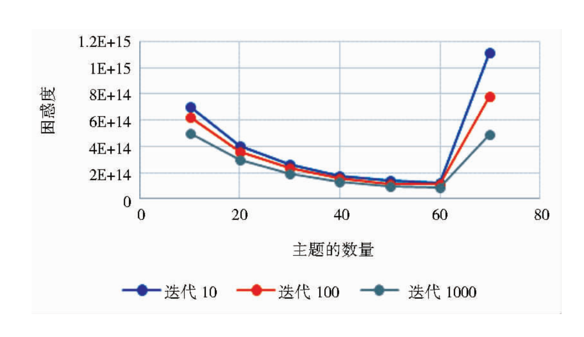


图 4 LDA 主题模型困惑度

计算疾病的主题分布,最后利用公式(3)计算两个疾病的定义相似度;基于空间向量的相似度计算过程,首先,导入相应的 TF-IDF 模型,然后,计算疾病相应部分的 TF-IDF,最后,利用 gensim 中的 similarities 计算相似度;基于集合的相似度计算过程是将两个疾病症状部分的特征词利用公式(4)进行计算。

4.3.2 结果分析

在疾病的综合语义相似度计算中,涉及到基于 DO 的疾病语义相似度和基于医学百科的的疾病语义相似度间的权重 α,β 对综合语义相似度计算的影响,在实验中设计了综合 1、综合 2、综合 3 三个不同的综合语义相似度,其中综合 1 语义相似度计算的权重设置:$\alpha=0.5$,$\beta=0.5$;综合 2 语义相似度计算的权重设置:$\alpha=0.6$,$\beta=0.4$;综合 3 语义相似度计算的权重设置:$\alpha=0.4$,$\beta=0.6$。在基于医学百科的疾病语义相似度计算中,涉及到概述、病因、症状、诊断和治疗方面的语义相似度间的权重 γ_1、γ_2、γ_3、γ_4、γ_5,由于概述部分是对疾病的表达,考虑其对疾病语义的影响程度高于其它部分,因此设置较高的权重;病因、症状、诊断和治疗方面四个方面分别从不同角度描述疾病,将这四个语义相似度对疾病语义相似度影响程度认为是均等的,因此四个权重相同;在实验中设计基于医学百科的疾病相似度定义为"综合 4",相关的权重设置为:$\gamma_1=0.4$,$\gamma_2=\gamma_3=\gamma_4=\gamma_5=0.15$。

利用本文提出的方法对 20 对疾病进行语义相似度实验,其中本体相似度、定义相似度、病因相似度、症状相似度、诊断相似度、和治疗相似度、以及百科语义相似度(综合 4)和综合语义相似度(综合 1、综合 2、综合 3)的实验结果见表 1。

在表 1 中,2 型糖尿病同糖尿病的语义相似度最高,由于 2 型糖尿病是糖尿病的一种类型,因此语义相似性也就最高;2 型糖尿病和糖尿病视网膜病变在本体中分别属于代谢类疾病和解剖类疾病,本体语义相似度为 0,实际上糖尿病视网膜病变是由糖尿病引起

的,因此两者之间存在相似性;同样,埃博拉出血热同病毒性脑膜炎和病毒性肺炎在定义描述和病因中都具有较高相似性,但是本体相似度为0。通过表1表明本文多维度的语义相似度计算能够反映出疾病的语义相似度。

为了分析疾病的本体相似度、定义相似度、病因相似度、症状相似度、诊断相似度、和治疗相似度、以及百科语义相似度(综合4)和综合语义相似度(综合1、综合2、综合3)等各种不同方法同人工判断的疾病语义相似度相关性,本文采用 Spearman 相关系数公式(10)和 Pearson 相关系数公式(11)分析各种不同相似度同人工相似度的相关性,实现结果如图5所示:

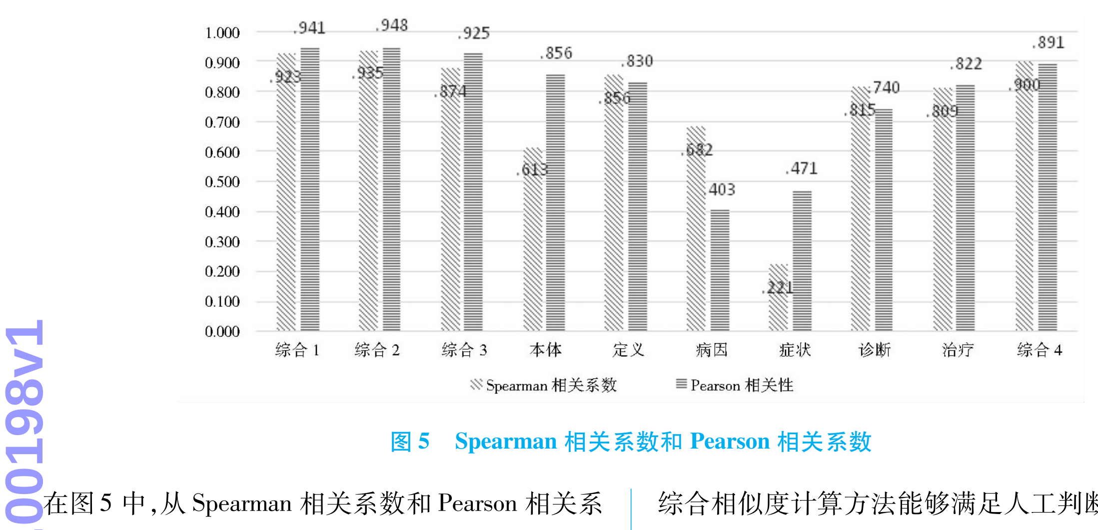


图5 Spearman 相关系数和 Pearson 相关系数

在图5中,从 Spearman 相关系数和 Pearson 相关系数整体看上,疾病的综合语义相似度(综合1、综合2、综合3)、本体相似度、定义相似度、病因相似度、症状相似度、诊断相似度、和治疗相似度、以及百科语义相似度(综合4)同人工判读的疾病与相似度具有较高相关性;从 Spearman 相关系数看,综合2语义相似度同人工判断的相关性最高(Spearman 相关系数为0.935),次之为综合1(Spearman 相关系数为0.923),再次为疾病百科语义相似度(综合4, Spearman 相关系数为0.900),综合3(Spearman 相关系数为0.874),症状相似度(Spearman 相关系数为0.221)最低;从 Pearson 相关系数看,同 Spearman 相关系数基本一致,综合2语义相关性(Pearson 相关系数为0.948)最高,次之为综合1(Pearson 相关系数为0.941),再次为综合3(Pearson 相关系数为0.925),症状相似度(Pearson 相关系数为0.471)最低。

同人工判读的疾病语义相似度实验表明:①综合疾病本体和疾病百科的疾病语义相似度计算结果优于分别利用疾病本体或疾病百科计算的语义相似度;②通过调整基于本体语义相似度和基于疾病百科的语义相似度的权重可以提高综合语义相似度的结果;③基于疾病百科的综合语义相似度明显优于将定义、病因、症状、诊断和治疗分别计算语义相似度;④综合1、综合2和综合3同人工判断相关系数表明,基于 DO 的疾病语义相似度相较于基于医学百科的疾病语义相似度的权重相对较高些。整体表明本文提出多维度的语义综合相似度计算方法能够满足人工判断疾病语义相似度要求,具有较好的效果。

5 结语

本文利用疾病本体和疾病百科设计了多维度的疾病语义相似度计算方案。针对医学百科全书中疾病概述部分,设计了基于 LDA 的语义相似度计算方法;针对疾病症状部分,设计了基于结合的语义相似度计算方法;针对病因、诊断和治疗部分,设计了基于空间向量的语义相似度计算方法;依据专家和实验将所有相似度其融合到一起,其中基于 DO 的疾病语义相似度和基于医学百科的疾病语义相似度间的权重在实验中分别采用三组不同值进行了探索分析,结果表明基于 DO 的疾病语义相似度相较于基于医学百科的疾病语义相似度的权重相对较高些。本文的方法既包含有疾病概念关系的相似度,又包括疾病的语义描述,从而实现多维度来衡量疾病的语义相似度,与单维度的语义相似度相比,具有较好的效果。下一步将继续研究基于医学百科语义相似度中融入医学知识,使疾病语义相似度的计算进一步优化和完善;进一步将本文提出的文本相似度运用到其他研究领域中。

参考文献:

[1] ILAKIVA P, SUMATHI M, KARTHIK S. A survey on semantic similarity between words in semantic Web[C]//International conference on radar, communication and computing. Tiruvannamalai: IEEE, 2012:213－216.

[2] 沙勇忠,史忠贤. 基于语义相似度的公共危机事件案例检索方法[J]. 情报资料工作, 2014(6):78-81.
[3] LIU L, YU Z. An improved knowledge push method based on semantic similarities[C]//Fourth international conference on multimedia information networking and security. Nanjing: IEEE, 2012: 378-380.
[4] 王道平,赵耀,刘涛. 敏捷供应链中知识服务检索的语义相似度问题研究[J]. 图书情报工作,2010,54(16):78-81.
[5] KULMANOV M, HOEHNDORF R. Evaluating the effect of annotation size on measures of semantic similarity[J]. Journal of biomedical semantics, 2017, 8(1):7.
[6] 李杰,初砚硕,程亮,等. 基于疾病本体的疾病相似性计算方法[J]. 生物化学与生物物理进展,2015, 42(2):115-122.
[7] NCBO BioPortal[EB/OL]. [2019-08-08]. https://bioportal.bioontology.org/.
[8] SCHRIML L, ARZE C, NADENDLA S, et al. Disease ontology: a backbone for disease semantic integration[J]. Nucleic acids research, 2012, 40(D1):D940-D946.
[9] 朱玲,杨峰, HE Y,等. 基本形形式化本体重要概念解析及对中医领域本体构建的提示[J]. 中国数字医学,2018,13(2):27-30,56.
[10] 陈云志. 肝炎本体构建及语义相似度研究[D]. 杭州:浙江大学, 2017.
[11] JORGE M. An overview of textual semantic similarity measures based on Web intelligence[J]. Artificial intelligence review, 2012, 42(4):935-943.
[12] 秦春秀,赵捧未,刘怀亮. 词语相似度计算研究[J]. 情报理论与实践,2007,30(1):105-108.
[13] SPAGNOLA S, LAGOZE C. Edge dependent pathway scoring for calculating semantic similarity in conceptnet[C]//Proceedings of the ninth international conference on computational semantics. Tilburg: Association for Computational Linguistics, 2011:385-389.
[14] CILIBRASI R, VITANYI M. The google similarity distance[J]. IEEE transactions on knowledge and data engineering, 2007, 19(3):370-383.
[15] 李峰,李芳. 中文词语语义相似度计算——基于《知网》2000[J]. 中文信息学报,2007(3):99-105.
[16] 刘杰,郭宇,汤世平,等. 基于《知网》2008 的词语相似度计算[J]. 小型微型计算机系统,2015,36(8):1728-1733.
[17] NGUYEN T, CONRAD S. A semantic similarity measure between nouns based on the structure of wordnet[C]//Proceedings of international conference on information integration and Web-based applications & services. Vienna: ACM, 2013:605-619.
[18] LIU X, ZHOU Y, ZHENG R. Measuring semantic similarity in wordnet[C]//International conference on machine learning and cybernetics. Hong Kong: IEEE, 2007: 3431-3435.
[19] 张军亮,朱学芳. 基于《农业大词典》的农业概念簇表示研究[J]. 情报科学,2013,31(7):15-17,22.
[20] 陈二静,姜恩波. 文本相似度计算方法研究综述[J]. 数据分析与知识发现,2017,1(6):1-11.
[21] AMINUL I, DIANA I. Semantic text similarity using corpus-based word similarity and string similarity[J/OL]. ACM Transactions on knowledge discovery from data, 2008,2(2):10. [2019-08-08]. http://www.researchgate.net/publication/220345072.
[22] CHEN Q, YAO L, YANG J. Short text classification based on LDA topic model[C]//International conference on audio, language and image processing. Shanghai: IEEE, 2016:749-753.
[23] FAROUK M. Sentence semantic similarity based on word embedding and WordNet[C]//13th international conference on computer engineering and systems. Cairo: IEEE, 2018:33-37.
[24] 李琳,李辉. 一种基于概念向量空间的文本相似度计算方法[J]. 数据分析与知识发现,2018,2(5):48-58.
[25] 詹志建,杨小平. 一种基于复杂网络的短文本语义相似度计算[J]. 中文信息学报,2016,30(4):71-80,89.
[26] 李慧. 词语相似度算法研究综述[J]. 现代情报,2015,35(4):172-177.
[27] BOLLEGALA D, ISHIZUKA M, MATSUO Y. Measuring semantic similarity between words using web search engines[C]//International conference on World Wide Web. Banff: ACM, 2007:757-766.
[28] ZHU G, IGLESIAS C. Computing semantic similarity of concepts in knowledge graphs[J]. IEEE transactions on knowledge and data engineering, 2017,29(1):72-85.
[29] RADAR, MILI H, BICHNELL E, et al. Development and application of a metric on semantic nets[J]. IEEE transaction on systems, man, and cybernetics. 1989,19(1): 17-30.
[30] BANU A, FATIMA S S, KHAN K U R. A new ontology-based semantic similarity measure for concepts subsumed by multiple super concepts[J]. International journal of Web applications, 2014, 6(1): 14-22.
[31] ZHU X, LI F, CHEN H, et al. An efficient path computing model for measuring semantic similarity using edge and density[J]. Knowledge and information systems, 2018, 55(1):79-111.
[32] 李文清,孙新,张常有,等. 一种本体概念的语义相似度计算方法[J]. 自动化学报,2012,38(2): 229-235.
[33] SAHNI L, SEHGAL A, KOCHAR A, et al, A novel approach to find semantic similarity measure between words[C]// 2nd international symposium on computational and business intelligence. New Delhi: IEEE, 2014:89-92.
[34] YANG Y, PING Y. An Ontology-based semantic similarity computation model[C]// IEEE international conference on big data and smart computing. Shanghai: IEEE, 2018:561-564.
[35] PESQUITA C, FARIA D, FALCÃO A O, et al. Semantic similarity in biomedical ontologies[J]. PLoS computational biology, 2009, 5(7):e1000443.
[36] DUTTA P, BASU S, KUNDU M. A new hybrid semantic similarity measure using information content and topological features of the Gene Ontology graph[C]//International conference on computer

communication and informatics. Coimbatore: IEEE, 2017:1 - 5.
[37] JEONG J, CHEN X. A new semantic functional similarity over gene ontology[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2015, 12(2):322 - 334.
[38] DUTTA P, BASU S, KUNDU M. Assessment of semantic similarity between proteins using information content and topological properties of the gene ontology graph[J]. IEEE/ACM transactions on computational biology & bioinformatics, 2018,15(3):839 - 849.
[39] AL-MUBAID H, NGUYEN H. Using MEDLINE as standard corpus for measuring semantic similarity in the biomedical domain [C]//Sixth IEEE international symposium on bioInformatics and bioEngineering. Arlington: IEEE, 2006: 315 - 318.
[40] 李文庆. 基于医学领域本体的语义相似度算法研究[D]. 太原:太原理工大学,2013.
[41] ZHANG J, ZHU X, ZHU G. Designing an automated FAQ answering system for farmers based on hybrid strategies[J]. Chinese journal of library and information science,2012,5(4):21 - 36.
[42] BLEI D, NG A, JORDAN M I, et al. Latent dirichlet allocation [J]. Journal of machine learning research, 2003,3(3): 993 - 1022.
[43] 何伟林,谢红玲,奉国和. 潜在狄利克雷分布模型研究综述[J]. 信息资源管理学报,2018,8(1):55 - 64.
[44] 刘铭,王晓龙,刘远超. 基于语义的高维数据聚类技术[J]. 电子学报,2009,37(5):925 - 929.
[45] Disease ontology[EB/OL]. [2019 - 08 - 08]. http://www.disease-ontology.org/.
[46] 百科名医[EB/OL]. [2019 - 08 - 08]. http://www.baikemy.com/.
[47] Python[EB/OL]. [2019 - 08 - 08]. http://www.python.org/.
[48] HanLP[EB/OL]. [2019 - 08 - 08]. http://hanlp.linrunsoft.com/.
[49] NumPy[EB/OL]. [2019 - 08 - 08]. http://www.numpy.org/.
[50] gensim:Topic modelling for humans[EB/OL]. [2019 - 08 - 08]. http://radimrehurek.com/gensim/.
[51] 周爱明. 图书情报领域实用多元统计[M]. 郑州:郑州大学出版社,2017.
[52] 关鹏,王曰芬. 科技情报分析中 LDA 主题模型最优主题数确定方法研究[J]. 现代图书情报技术,2016(9):42 - 50.

Research on Semantic Similarity of Diseases Based on Multidimensional

Zhang Junliang[1,2,3]
[1] School of Management, Xinxiang Medical University, Xinxiang 453003
[2] Center for Health Information Resources, Xinxiang Medical University, Xinxiang 453003
[3] Institutes of Health Central Plain, Xinxiang 453003

Abstract: [**Purpose/significance**] Aiming at different expression of disease knowledge, this paper proposes a comprehensive semantic similarity calculation scheme that integrates multi-dimension of disease. [**Method/process**] On the basis of integrating the characteristics of disease ontology and Medical Encyclopedia, the comprehensive semantic similarity, which consists of semantic similarity based on disease ontology and disease semantic similarity based on medical encyclopedia, was built. Semantic similarity of diseases based on medical encyclopedia was calculated by LDA, set theory and vector space model. [**Result/conclusion**] The results show that the proposed method can effectively reflect the semantic similarity of diseases. The comprehensive semantic similarity calculation scheme offers helpful reference for further research.

Keywords: semantic similarity disease ontology disease encyclopedia similarity measure